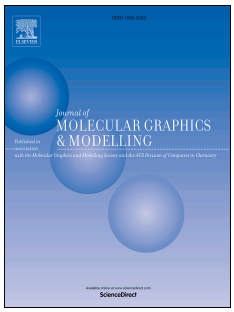
Accepted Manuscript



Shape complementarity at protein interfaces via global docking optimisation

Gareth Williams

PII: S1093-3263(18)30278-X

DOI: 10.1016/j.jmgm.2018.06.011

Reference: JMG 7187

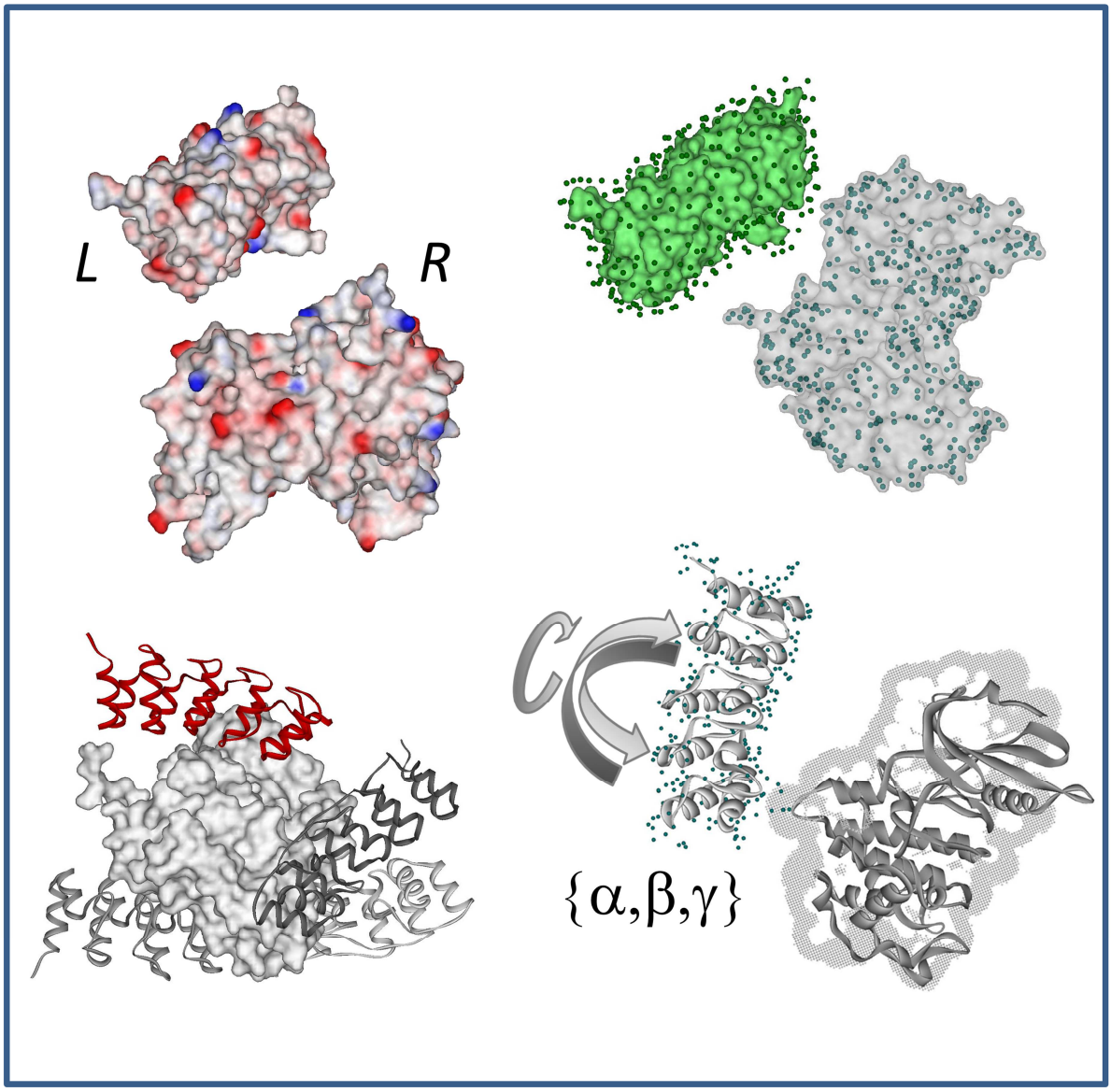
To appear in: *Journal of Molecular Graphics and Modelling*

Received Date: 19 April 2018

Revised Date: 11 June 2018

Accepted Date: 12 June 2018

Please cite this article as: G. Williams, Shape complementarity at protein interfaces via global docking optimisation, *Journal of Molecular Graphics and Modelling* (2018), doi: 10.1016/j.jmgm.2018.06.011.



Shape complementarity at protein interfaces via global docking optimisation

Gareth Williams

Bioinformatics, Wolfson CARD, The Wolfson Wing, Hodgkin Building, King's College London, SE1 1UL (gareth.2.williams@kcl.ac.uk).

Abstract

Protein complexes are characterised by shape complementarity at the interface. Here we present a simple fast global shape fitting algorithm to investigate the extent to which interfaces are global minima of complementarity. The algorithm is applied to a varied set of hetero and homo complexes and complexes between complexes showing that over 90% of large interfaces are global maxima in the space of shape complementarity.

Keywords

Protein structure
Protein-protein docking
Shape complementarity

Introduction

Protein interaction interfaces are characterised by a high degree of shape complementarity [1, 2]. Lawrence and Coleman [1] gave a quantitative score to the goodness of fit at a protein interface through a dot product of surface vectors corresponding to proximal atoms. This technique served as the basis for a concise Ramachandran-like 2D plot representation of protein interfaces combining shape and electrostatic complementarity (Sc and EC)[3]. Machine learning implementations based on Sc, EC and interface size have been successful in filtering out true native-like docked conformations from a dataset of possible poses [4]. The extent to which crystallographic complex solutions deviate from interface complementarity can also be scored and visualised via a small radius probe, implemented with the Molprobity web tool [5]. Molprobity has served as an effective tool for crystal structure optimisation.

Global shape complementarity docking algorithms such as GRAMM have shown that native-like docked complexes emerge as those with optimal surface overlap at various levels of coarse granularity [6]. Such exhaustive searches are over a six dimensional space with an additional scoring of the interface and are usually speeded up with techniques such as the fast Fourier transform (FFT). This methodology has been extended in FTDock [7] to include an electrostatic filter to separate high complementarity poses. The FFT overlap calculation speed up has been the basis of ClusPro [8], where high overlap conformations filtered based on statistical potentials and then clustered, and DOT [9]. Alternatively, protein structures have been approximated with a low order spherical polar Fourier expansion with a resulting relatively fast ‘surface skin’ correlation calculation [10]. Non-exhaustive techniques have also been developed based on defining surface features according to local convexity, concavity and flatness. In this PatchDock approach only conformations with matching patches are scored for complementarity [11]. In a further level of abstraction, the binding site shape has been shown to be describable with the first few terms of a Zernike 3D shape descriptor polynomial leading to a relatively rapid complementarity calculation [12]. The methodology presented here is based on an initial coarse grained exploration of conformations pivoted on surface atom surface point pairs followed by a fine scale analysis of limited set of putative binding conformations. The methodology recovers 92% of a mixed set of protein complex conformations.

Results

The algorithm was tested on a mixed set of 278 protein complexes. These consisted of 87 homo-oligomer complexes [13] and 191 transient hetero-oligomer complexes [14], with 65 of these complexes involving more than two proteins. The complexes are non-redundant in the sense that they don’t share sequence homology at the interface. The results described below are not sensitive to the particular set of complexes examined. The first stage coarse grained docking is sufficient to identify 61% of the complexes as optimal overlap conformations. A native-like docked conformation is called optimal if it comes in the top four ranked conformations according to the given scoring system. This percentage rises to 81% with the fine grained optimisation. The likelihood of a binding conformation being an optimal in

shape complementarity increases with the size of the binding interface, see Table 1. In particular, for complexes with a number of atom contacts (4Å proximity) greater than 400 92% are complementarity optimal. This constitutes 80% of the complexes. It appears that measuring the size of the interaction through the amount of buried accessible surface area (ASA) is worse at segregating optimal from non-optimal complementarity, see Figure 3. Here, only 84% of the top 80% ASA complexes are complementarity optimal. An example docking run is shown in Figure 4. Here, an antibody light/heavy chain pair is docked with its target (von Willebrand factor pdb accession 1fe8). The 'ligand' scores highly at multiple sites on the antibody, with two conformations (ranked 1 and 3) in the top five aligning with the native structure.

The atom type content of the interface is invisible to the analysis so far. A simple way to introduce atom type content in interface description is through a vector over a relatively small set of properties. In particular, five types of atom are considered: neutral, donor, acceptor, positive, negative. Thus each docking pose is associated with a docking matrix. A simple linear model can then be used to maximise the score associated with the native-like docking conformation. In particular, collecting 190 poses with corresponding RMSDs a linear model predicts slightly more native-like docked conformations, 86% and 96% of large interfaces. The relative contribution of the various atom pairings to the native-like docked interface relative to other complementary interfaces is shown in Table 2. The main contribution by virtue of being the dominant atop type comes from neutral pairings. As expected, opposite charge pairings and pairings between acceptor/donor and positive/negative atoms also contribute positively. The biggest effect is on the smaller interfaces where the native-like docked conformations now constitute 49% of the high scoring complementarity poses as opposed to 40% without atom type information.

Amino acid preferences in both inter- and intra-protein interaction have been the subject of much research. The propensities for internal contacts have been developed into statistical potentials [15-17] that have been employed in protein folding simulations [18]. Propensities of amino acid types at protein interfaces have been effectively deployed as supplements to docking scores [19-23]. With this in mind it is of interest to investigate to what extent high complementarity interfaces segregate between native and non-native on the basis of amino acid content. Amino acid type data can be introduced in the same way as atom type data. However, a linear model fit based on the amino acid content and contact number at the interface only results in a moderate improvement in predictability from 81% to 83%. The beta factors for the amino acid contribution have a small but significant correlation with the probabilities associated with the Miyazawa Jernigan amino acid interaction energies (Pearson correlation -0.25 Zscore 3.69).

Methods

Each protein or protein complex to be docked is arbitrarily separated into a 'ligand' and 'receptor' pair. The 'ligand' is mobile and the 'receptor' is stationary. The protein surface of the 'ligand' is reduced to a set of points separated by 3.5Å from each other and the protein heavy atoms. The 'receptor' surface points will anchor 'ligand' surface atoms hence the

separation of the surface point from the 'receptor' heavy atoms is informed by nearest neighbour heavy atom separations observed in protein complexes. In the interest of computational speed surface points are spread at separations of 3.5Å. Heavy atoms proximal to surface points are called surface atoms. A 0.5Å cubic lattice spanning the 'receptor' is defined with lattice points either assigned to the interior of the 'receptor' or the surface according to proximity. Specifically, initially a surface cloud is defined extending 5Å from each surface atom and then a cloud corresponding to the protein interior is defined extending 2Å from each heavy atom. Thus the surface cloud has a thickness of 3Å and has an average distance from the protein heavy atoms of 3.5Å, which is the observed interaction distance between heavy atoms. The algorithm proceeds by locking each surface atom of the 'ligand' to each surface point of the receptor and generating conformations through Euler rotations. Each conformation is first assessed for clashes with the 'receptor' interior cloud. This is implemented in a two stage process by first looking for clashes of the nearest neighbour ligand heavy atoms to the anchored ligand atom and then looking for clashes across the whole ligand surface atom set. Conformations are scored by the number of surface atoms occupying the surface cloud of the receptor. The 9,000 top ranked conformations are then scored for the number of 'ligand'-'receptor' heavy atom contacts, defined by proximity of less than 4Å. Specifically, the docking score is defined as the geometric average of the number of atoms at the interface and the number of contacts made by the interface atoms, $s = \left(\frac{B_1}{N_1} + \frac{B_2}{N_2}\right)(N_1 + N_2)$, where $B_{1,2}$ are the numbers of interaction contacts made by the atoms of the two proteins and $N_{1,2}$ are the numbers of atoms of each protein contributing to the interface. The final predicted conformation has the highest number of 'ligand' receptor contacts.

The algorithm is exhaustive, but the initial global conformational search is coarse grained over a limited number of surface point and atoms. Rotational orientations are generated through a relatively coarse solid angle increment of 0.01π giving a total of 420 configurations. The initial phase involves a number of moves scaling as the second power of the ligand surface atom number and a single power of the receptor surface point number. For computational speed the 'ligand' is taken to be the smallest of the pair to be docked. A typical complex involves $2000 \times N_1 \times N_2$ moves in the initial phase, where $N_{1,2}$ are the number of residues in the proteins to be docked. The final optimisation is a trivial time increment over the initial phase.

The linear fitting is according to:

$$y_i \sim y_i' = \sum_{nm} \beta_{nm} x_i^{nm} + c,$$

where *y* is a categorical call on whether the RMSD of the docking pose is within 3Å of the native conformation ($y = 1,0$ according to whether the pose is native/non-native), *x* are the numbers of contacts of the given types at the interface and *c* a constant. For each protein pair the poses are now ranked based on the β factors and the success of recovering a native-like docked pose is measured by the rank of the native-like docked fold amongst alternatives. The relative contribution of the different atom type pairings can be seen by fitting to the

normalised contact matrix In particular, the elements are normalised over the set of interfaces i.e. $x_i^{nm} \rightarrow \frac{x_i^{nm} - \langle x^{nm} \rangle}{\sigma(x^{nm})}$, where the brackets refer to the average over the interfaces and σ is the standard deviation.

Conclusions

A simple shape complementarity algorithm has been used to show that protein interactions above a certain size correspond to maxima in the space of interface overlaps. In particular, 92% of complexes with relatively large interfaces are optimal for shape complementarity. Only a relatively small fraction of small interfaces are optimal for complementarity. However, these interfaces are characterised by pairings of specific atom types. In particular, when atom type information is introduced in the simplest form of five basic categories (neutral, donor, acceptor, positive, negative) then the majority of small interfaces are also recovered through optimisation. The contribution of this rudimentary electrostatic information only makes a marginal impact on the correlation with native-like docked conformations and it is difficult to see how more subtle statistical potentials would be more critical. Further, introducing amino acid type content at the interface has less of an impact on segregation of native and non-native complexes than atom type content. The algorithm presented here overcomes the problems of conformation space size and intricate overlap calculation by introducing a two stage docking process. In the first stage a relatively small set of 'ligand' surface points are pivoted on a relatively small set of 'receptor' surface associated atoms and the possible arrangements sampled with a coarse grained Euler angle set. In the second stage a set of high scoring poses from the first phase are scored for a full atom surface overlap. The technique is different to those published and does not rely on FFT or any surface shape abstractions. In conclusion, the docking presented here, in line with most published methods, is for rigid bodies where the backbone and side chains are fixed as given in the crystal derived coordinates. Incorporating residue flexibility in docking is a much more complex process involving the generation of multiple conformers and then combining soft-docking with refinement, see Andrusier et al for a review of these methods [24]. However, protein flexibility is outside the scope of the present study. Rather, the question of the extent of native-like protein-protein interface complementarity in the ensemble high complementarity binding conformations is addressed.

References

[1] Lawrence, M.C., Colman, P.M. Shape complementarity at protein/protein interfaces. J Mol Biol. 1993, 234, 946-50.
[2] Jones, S., Thornton, J.M. Protein-protein interactions: a review of protein dimer structures. Progress in biophysics and molecular biology. 1995, 63, 31-65.
[3] Basu, S. CPdock: the complementarity plot for docking of proteins: implementing multi-dielectric continuum electrostatics. Journal of molecular modeling. 2017, 24, 8.
[4] Basu, S., Wallner, B. Finding correct protein-protein docking models using ProQDock. Bioinformatics. 2016, 32, i262-i70.

[5] Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res. 2007, 35, W375-83.
[6] Vakser, I.A. Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. Proteins. 1997, Suppl 1, 226-30.
[7] Gabb, H.A., Jackson, R.M., Sternberg, M.J. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol. 1997, 272, 106-20.
[8] Comeau, S.R., Gatchell, D.W., Vajda, S., Camacho, C.J. ClusPro: a fully automated algorithm for protein-protein docking. Nucleic Acids Res. 2004, 32, W96-9.
[9] Roberts, V.A., Thompson, E.E., Pique, M.E., Perez, M.S., Ten Eyck, L.F. DOT2: Macromolecular docking with improved biophysical models. Journal of computational chemistry. 2013, 34, 1743-58.
[10] Ritchie, D.W., Kemp, G.J. Protein docking using spherical polar Fourier correlations. Proteins. 2000, 39, 178-94.
[11] Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., Wolfson, H.J. PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res. 2005, 33, W363-7.
[12] Venkatraman, V., Yang, Y.D., Sael, L., Kihara, D. Protein-protein docking using region-based 3D Zernike descriptors. BMC Bioinformatics. 2009, 10, 407.
[13] Bahadur, R.P., Chakrabarti, P., Rodier, F., Janin, J. Dissecting subunit interfaces in homodimeric proteins. Proteins. 2003, 53, 708-19.
[14] Mintseris, J., Weng, Z. Atomic contact vectors in protein-protein recognition. Proteins. 2003, 53, 629-39.
[15] Miyazawa, S., Jernigan, R.L. Estimation of effective interresidue contact energies from protein crystal structures- Quasi-chemical approximation. . Macromolecules. 1985, 18, 534-52.
[16] Miyazawa, S., Jernigan, R.L. An empirical energy potential with a reference state for protein fold and sequence recognition. Proteins. 1999, 36, 357-69.
[17] Miyazawa, S., Jernigan, R.L. How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins? J Chem Phys. 2005, 122, 024901.
[18] Jernigan, R.L., Bahar, I. Structure-derived potentials and protein simulations. Curr Opin Struct Biol. 1996, 6, 195-209.
[19] Moont, G., Gabb, H.A., Sternberg, M.J. Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins. 1999, 35, 364-73.
[20] Lu, H., Lu, L., Skolnick, J. Development of unified statistical potentials describing protein-protein interactions. Biophys J. 2003, 84, 1895-901.
[21] Huang, S.Y., Zou, X. An iterative knowledge-based scoring function for protein-protein recognition. Proteins. 2008, 72, 557-79.
[22] Liu, S., Vakser, I.A. DECK: Distance and environment-dependent, coarse-grained, knowledge-based potentials for protein-protein docking. BMC Bioinformatics. 2011, 12, 280.
[23] Nadalin, F., Carbone, A. Protein-protein interaction specificity is captured by contact preferences and interface composition. Bioinformatics. 2018, 34, 459-68.
[24] Andrusier, N., Mashiach, E., Nussinov, R., Wolfson, H.J. Principles of flexible protein-protein docking. Proteins. 2008, 73, 271-89.

Legends

Figure 1. Surface points and surface associated atoms in a protein complex. A complex of a chemokine ligand 5 homo-dimer (pdb 1b3a) with the surface points of the 'ligand' shown in light grey and the surface associated atoms of the 'receptor' shown in grey. The proteins are shown as space filled structures with a 1.5 A probe, with the receptor transparent.

The surface points lie outside and the surface associated atoms lie within the surface. To the right the surface points in proximity with the surface associate atoms are shown demonstrating that these pairs can facilitate pivot points for docking poses.

Figure2. Surface atoms imbedded in the lattice surface cloud. A complex of a chemokine ligand 5 homo-dimer (pdb 1b3a) is shown with a slice of chain A surface associated atoms in light grey with the lattice surface cloud (grey mesh). Chain B is shown transparent space filled (1.5A probe radius) together with those surface associated atoms that are imbedded in the surface cloud of chain B, dark grey. Docking is initially scored by the number of surface associated ligand atoms buried in the receptor surface cloud.

Figure3. Native-like docked complexes segregate based on interface size. Plots of the fraction of top docking configurations that agree with the crystal data versus the interface size. Solid lines refer to the full optimisation and the broken lines to the initial coarse-grained stage. The interface size can either be measured by the number of proximal pairs, shown left, or the amount of buried surface area, shown right. It is clear that contact number more effectively predicts complex recovery.

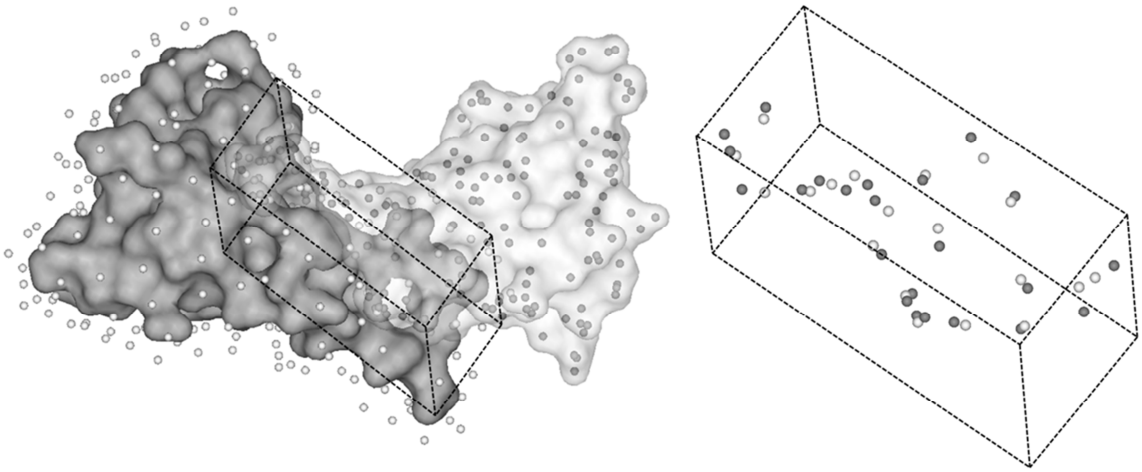
Figure 4. Example docking poses with high complementarity. The top five ranked poses by complementarity of the von Willebrand factor with the antibody light and heavy chain complex (structures from 1fe8). The top scoring pose is shown together with the true conformation on the right. The RMSD between the ligands is 1.21A. The third ranked pose is slightly closer to the true conformation, with an RMSD of 0.85A.

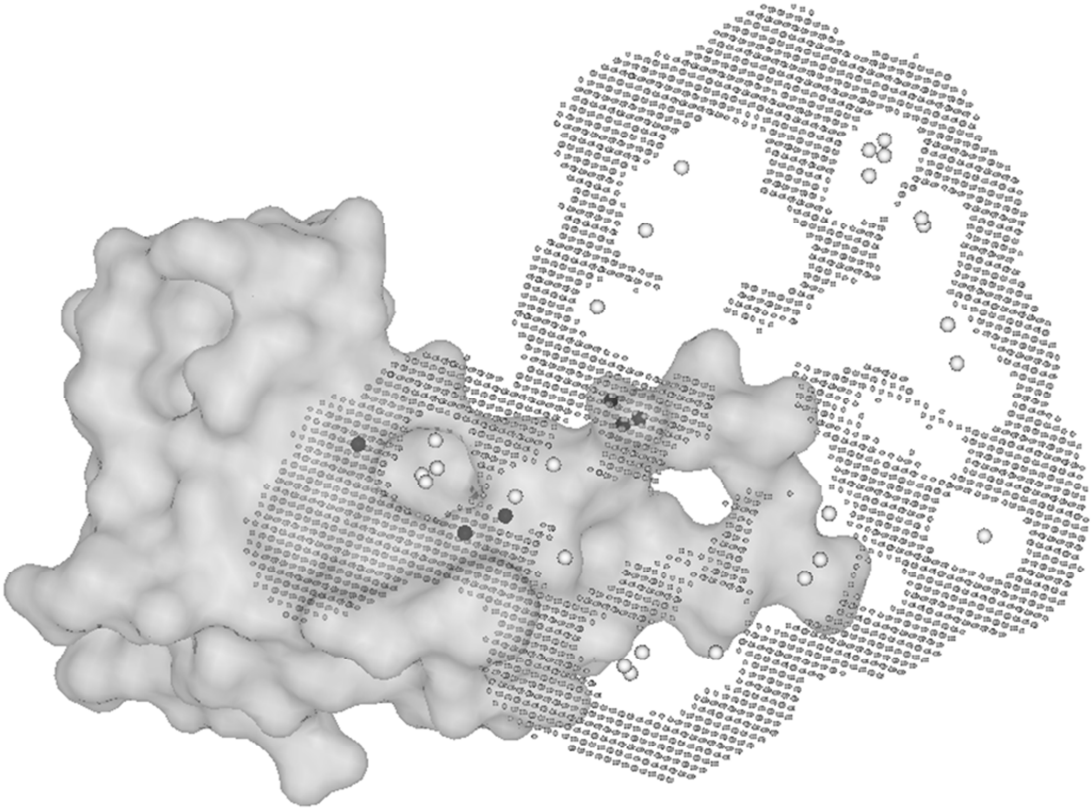
Table1: The protein complexes used in the analysis. The protein data bank identifier is given together with the two chain identifiers. The colon separates the single or multiple domains to be docked.

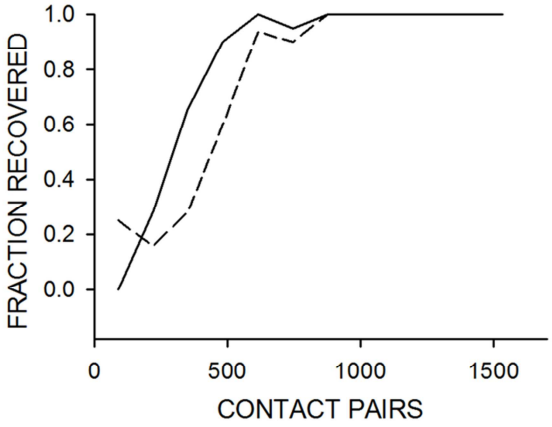
Table 2: Linear model fit relative contribution of different atom type pairings at native-like docked interfaces. The relative contribution of the various atom type pairings at the native interface relative to non-native high complementarity interfaces as measured by the beta factors in the linear model fit. Apart from the dominant neutral contribution there is a high contribution from opposite charge, donor:negative and acceptor:positive pairings.

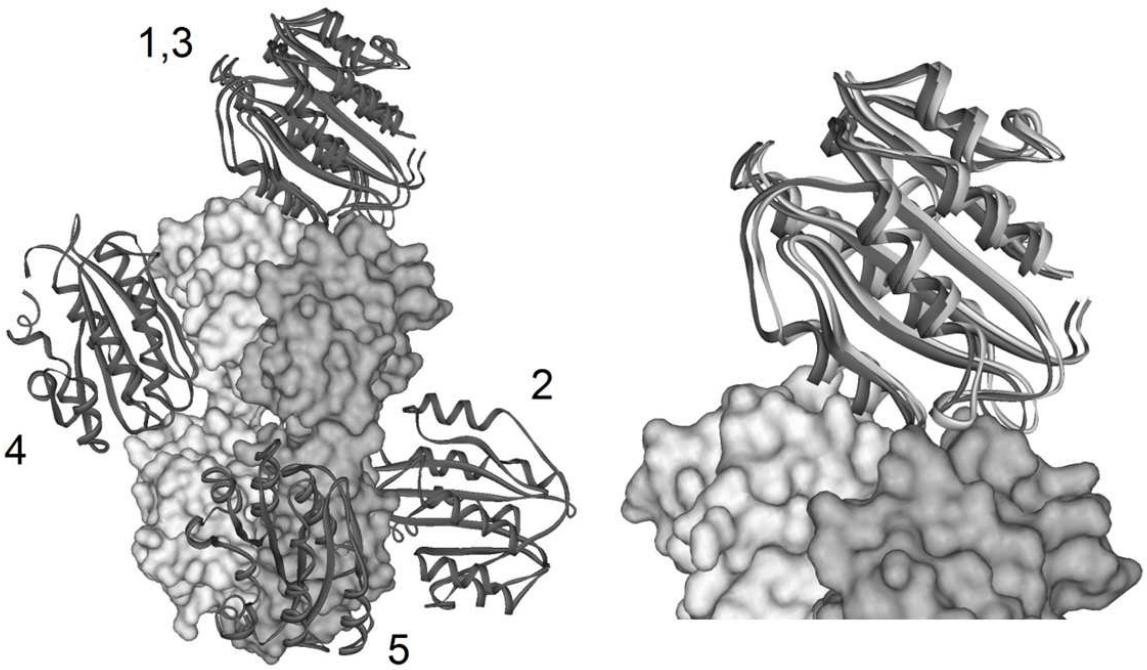
12as B:A	1bkd R:S	1dfj I:E	1fbi HL:X	1i7w A:B	1ncc N:HL	1ugh I:E
1a2k A:D	1bkp B:A	1dhk A:B	1fc2 C:D	1i8l A:C	1nfd EF:AB	1vfr B:A
1a2y AB:C	1blx A:B	1dkg AB:D	1fe8 A:HL	1iar A:B	1nrn HL:R	1vok B:A
1a4i B:A	1bmd B:A	1dor B:A	1fip B:A	1ib1 AB:E	1nse B:A	1vrk A:B
1a4u B:A	1bml A:C	1dpg B:A	1fj1 AB:F	1ibr A:B	1nsn HL:S	1wej HL:F
1a4y A:B	1bp3 A:B	1dpj A:B	1fle E:I	1icf AB:I	1nsy B:A	1wq1 G:R
1aa7 B:A	1bqq T:M	1dqj AB:C	1flt VW:Y	1icw B:A	1osp HL:O	1wtl B:A
1acb E:I	1brw B:A	1dqs B:A	1fns HL:A	1ihs HL:I	1pgt B:A	1www VW:Y
1ad3 B:A	1bsl B:A	1ds6 A:B	1foe A:B	1im3 A:D	1ppf I:E	1xso B:A
1ade B:A	1bsr B:A	1dtd A:B	1fq1 A:B	1im9 A:D	1pre B:A	1ycs A:B
1adq A:HL	1buh A:B	1du3 DF:A	1fqk A:B	1imb B:A	1qa9 A:B	1zbd A:B
1afw B:A	1bvn P:T	1dx5 AM:I	1fqv A:B	1iqd AB:C	1qav A:B	2arc B:A
1agr A:E	1bxg B:A	1dxg B:A	1fro B:A	1ira X:Y	1qfh B:A	2btc I:E
1aip A:CD	1bxk B:A	1dzb X:A	1fsk A:BC	1isa B:A	1qfu AB:HL	2btf P:A
1ajs B:A	1bzq A:L	1e0o A:B	1fyh A:B	1itb A:B	1qhi B:A	2ccy B:A
1ak4 A:D	1c1y A:B	1e44 A:B	1g3n A:C	1ivy B:A	1qkz A:HL	2hdh B:A
1aq6 B:A	1c4z A:D	1e6j HL:P	1g4u R:S	1j7v R:L	1qmz A:B	2hmi B:CD
1atn A:D	1ca0 BC:D	1e96 A:B	1g4y R:B	1jdp H:A	1qo0 A:DE	2jel HL:P
1auo B:A	1cd9 A:B	1eai A:C	1g73 A:D	1jhl HL:A	1qo3 A:CD	2lig B:A
1ava A:C	1cdc A:B	1eay A:C	1g9m C:G	1jiw P:I	1qr2 B:A	2mcg 2:1
1avg HL:I	1cdk A:I	1ebh B:A	1g9m G:HL	1jlt A:B	1r2f B:A	2nac B:A
1avw A:B	1cdm A:B	1ebp A:CD	1gcq B:C	1jma A:B	1reg Y:X	2ohx B:A
1avz B:C	1cg2 C:B	1efu A:B	1gh6 A:B	1jps HL:T	1rfb B:A	2pcc A:B
1axi A:B	1chm B:A	1egj A:HL	1gl0 I:E	1jrh HL:I	1rlb ABCD:F	2sic E:I
1ay7 A:B	1cho E:F	1eja A:B	1gl4 A:B	1jtd A:B	1rrp A:B	2spc B:A
1azz A:CD	1cic AB:CD	1emv A:B	1got A:B	1jtg A:B	1sbb A:B	2vir AB:C
1b2s A:D	1clv I:A	1eo8 AB:HL	1hcf AB:Y	1jtp A:L	1ses B:A	3bth I:E
1b3a B:A	1cmb B:A	1es7 AC:D	1he1 A:C	1k4c AB:C	1sgp I:E	3dap B:A
1b5e B:A	1cmx A:B	1euv A:B	1hez AB:E	1k90 A:D	1slt B:A	3sdh B:A
1b67 B:A	1cn4 AB:C	1ev2 A:E	1hia AB:I	1k9o E:I	1slu A:B	3ygs P:C
1b6c A:B	1cnz B:A	1f02 I:T	1hjr D:B	1kac A:B	1smn B:A	4cha C:B
1b8a B:A	1coz B:A	1f34 A:B	1hss B:A	1kba B:A	1smt B:A	4htc HL:I
1b8j B:A	1cse I:E	1f3v A:B	1hx1 A:B	1kcg AB:C	1sox B:A	4kbp C:B
1bbh B:A	1cxz A:B	1f51 AB:E	1hxp B:A	1kig HL:I	1stf I:E	4sgb I:E
1bd0 B:A	1d2z A:B	1f5q A:B	1hyr AB:C	1lfd AC:B	1tbr HL:R	5rub B:A
1bdj A:B	1d5m A:C	1f60 A:B	1i1r A:B	1lpb A:B	1tc1 B:A	7cei A:B
1biq B:A	1d6r I:A	1f7z A:I	1i2m A:B	1lyn B:A	1tmq A:B	8prk B:A
1bis B:A	1daa B:A	1f93 AB:EF	1i4d AB:D	1m6p B:A	1tnr A:R	9wga B:A
1bj1 HL:W	1dee CD:G	1fak HL:T	1i4o A:C	1mkb B:A	1toc R:AB	
1bjw B:A	1dev A:B	1fak I:HL	1i5k A:C	1mlc AB:E	1tx4 A:B	

	NEUTRAL	DONOR	ACCEPTOR	POSITIVE	NEGATIVE
NEUTRAL	8.76				
DONOR	-2.76	2.24			
ACCEPTOR	1.43	1.91	-0.52		
POSITIVE	-0.54	-1.73	2.86	-0.31	
NEGATIVE	-0.63	4.19	-3.81	5.86	-1.41









- A fast simple algorithm has been developed to explore the relative surface complementarity at native protein interfaces of native conformations.
- Native protein complexes are shown to emerge as interfaces ranked highest by complementarity.
- Atom type information encoding polarity and charge are shown to improve native structure recovery rates.
- Amino acid content at interfaces has a marginal effect on the segregation of native and non-native complexes.